

Kaufman - Trading Systems & Methods - Chap03 12-21

Chap03 12-21

VISIT...

LANZAROTE
Caliente.COM

Table 34 shows the relationship using selected values of x . Notice that this relationship is essentially linear due to the small secondorder coefficient. It is not necessary to solve both the linear and curvilinear models, because the secondorder equation can be used in the linear form when this situation occurs.

Transforming Nonlinear to Linear

Two curves that are often used to forecast prices are logarithmic (power) and exponential relationships (see Figure 3'). The exponential, curving up, is used to scale price data that become more volatile at higher levels. Each of these forms can be solved with unique equations,' however, both can be easily transformed into linear relationships and solved using the method of least squares. This will allow you to fool the computer into solving a nonlinear problem using a linear regression tool.

	Logarithmic	Exponential
General form	$y = ax^b$	$y = ae^{bx}$
Transformed	$\ln y = \ln a + b \ln x$	$\ln y = \ln a + bx$
Linear form	$y = a + bx$ where $y = \ln y$ $a = \ln a$ $x = \ln x$	$y = a + bx$ where $y = \ln y$ $a = \ln a$

The significant difference in the transformations is that the value of x is not scaled for the exponential. Taking the original data and performing the appropriate natural log functions, In results in the linear form that can be solved using least squares. Selected results from the computer program in Appendix 2 are shown in Table 35.

EVALUATION OF 2VARIABLE TECHNIQUES

Of the three curvefitting techniques, the curvilinear and exponential results are very similar. both curving upward and passing through the main cluster of data points at about the same incline. The log approximation curves downward after passing through the main group of data points at about the same place as the other approximations. To evaluate objectively whether any of the nonlinear methods are a better fit than the linear approximation. find the standard deviation of the errors, which gives a statistical measurement of how close the fitted line comes to the original data points. The results show that the curvilinear is best: the logarithmic, which curves downward, is noticeably the worst (see **Figure 38**).

In the case of the cornsoybean relationship, the model with the smallest variance makes the most sense. Prices will usually remain within the range tested. and the realities of the productionprice relationship support the selection of the linear (or curvilinear) values,

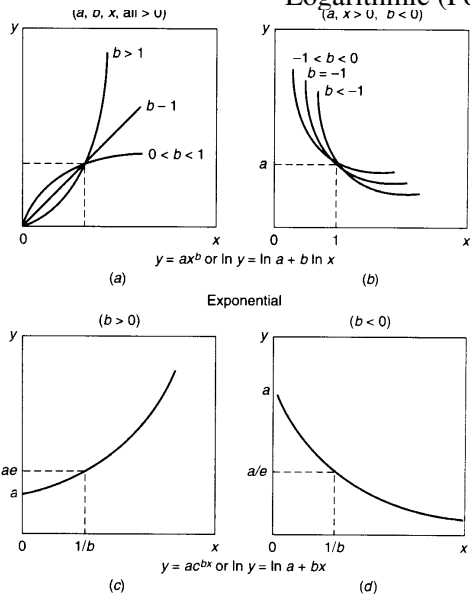
The use of regression analysis for forecasting the price of soybeans alone has a different conclusion. Although 27 years were used, the last 10 showed a noticeable increase in soybean prices. This rising pattern is best fit by, the curvilinear and exponential models. However, forecasts using these formulas show prices continuing to rise at an increasing rate. Had inflation maintained its doubledigit rate. these forecasts would still lead to unrealistically high prices. The logarithmic model. which leveled off after the rise. turned out to

TABLE 3-4 Curvilinear Values for Corn and Soybeans

Soybeans	x	1.00	2.00	3.00	4.00	5.00	6.00	7.00
Corn	y	.63	.96	1.29	1.63	1.97	2.31	2.66

49

FIGURE 37 Logarithmic and exponential.



Source: Cuthbert Daniel and Fred S.Wood, *Fitting Equations to Data: Computer Analysis of Multifactor* Dow, 2nd ed. (New York: John Wiley & Sons, 1980, pp. 20,21). Reprinted with permission .

be the best for the actual situation. This shows that the problem of forecasting is more complex than this naive solution. The logarithmic model, which showed the worst statistical results, provided the best forecast. Major factors that cause significant price shifts, such as interest rates, inflation, and the value of the U.S. dollar, must be monitored carefully. The model must be reestimated whenever these factors change. The section "Approximations" will discuss this in more detail.

Direct Relationships

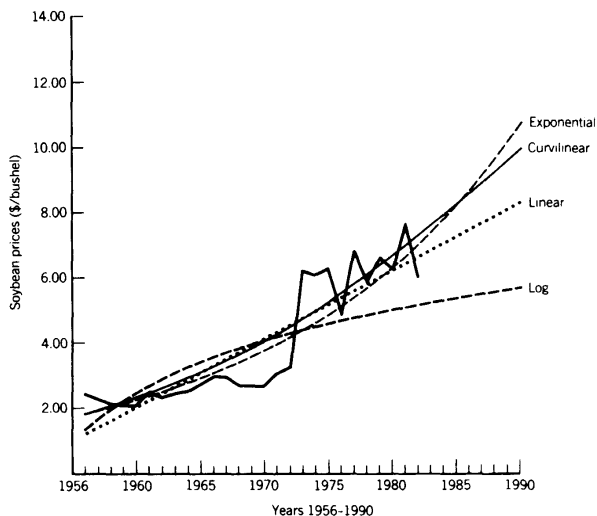
The price at which a market trades is limited by the cost of production and dependent upon the prices of other markets that provide a substitute product. Arbitrage is based on

TABLE 3-5 Log and Exponential Values for Corn and Soybeans

Soybeans	x	1.00	2.00	3.00	4.00	5.00	6.00	7.00
Corn (log)	y	.54	.94	1.30	1.64	1.96	2.26	2.56
(exp)	y	.84	1.02	1.24	1.50	1.83	2.22	2.70

50

FIGURE 38 Leastsquares approximation for soybeans using linear, curvilinear, logarithmic, and exponential models.



the ability to substitute one product for another after transaction costs, which can include carrying charges, shipping, inspection, and commissions. These relationships are carefully monitored and arbitrage opportunities are quickly acted upon by traders. This process keeps futures and cash prices together and prevents the price of gold in New York, London, and Hong Kong from drifting apart any further than the cost of buying gold in one location and delivering it to another. For interest rate markets, strips prevent the large pool of 3month vehicles, notes, and bonds from offering widely different returns for the same maturity. The Interbank market provides the same stability for foreign exchange markets, while in agriculture the soybean crush and other processing margins do not stay outofline for long. The following nonfinancial markets have close relationships that can be found using regression analysis.

Products Related	Reason
All crops	Relative income for farmers
Grains	Protein substitution for feed
Livestock to feedgrains	Feed, cost of production
Livestock	Food substitution
Sugar and corn	Sweetener substitution
Hogs and pork bellies	Product dependency
Silver, gold, and platinum	Investor's inflation hedge
Interest rates and financials	The stock market and foreign exchange are strongly influenced by changes in interest rates.

MULTIVARIATE APPROXIMATIONS

Regression analysis is most often used in complex economic models to find the combination of two or more independent variables that best explain or forecast prices. A simple application of annual production and distribution of soybeans will determine whether these factors are significant in determination of soybean prices. Because the demand for soybeans and its products is complex, a high correlation should not be expected between the data. However, there is no way of knowing how much impact other factors have on the prices over a long term.

Using the method of least squares, which was employed for a simple linear regression, the new equation is

$$y = a + bx_1 + cx_2$$

where y is the resulting price
 x_1 is the total production (supply)
 x_2 is the total distribution (demand)
 $a, b,$ and c are constants to be calculated

As in the linear approximation, the solution to this problem will be found by minimize ing the sum of the squares of the errors at each point

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Substituting $\hat{y}$ from the previous equation,

$$S = \sum_{i=1}^n (y_i - (a + bx_1 + cx_2))^2$$

The solution to the multivariate problem of two ind requires the following three least-squares equations:

$$\begin{aligned} an + b \sum x_1 + c \sum x_2 &= \sum y \\ a \sum x_1 + b \sum x_1^2 + c \sum x_1x_2 &= \sum x_1y \\ a \sum x_2 + b \sum x_1x_2 + c \sum x_2^2 &= \sum x_2y \end{aligned}$$

The procedure for solving the three simultaneous equations is the same as the curvilinear method of coefficient elimination. The sums are calculated in Table 36, then substituted into the last three equations:

$$\begin{aligned} 12.00a + 13.546b + 12.582c &= 44.4 \\ 13.55a + 15.980b + 14.75c &= 53.28 \\ 12.58a + 14.75b + 13.86c &= 49.264 \end{aligned}$$

The coefficient matrix solution is4

$$\begin{pmatrix} 12.00 & 13.55 & 12.58 & 44.40 \\ 13.55 & 15.98 & 14.74 & 53.28 \\ 12.58 & 14.74 & 13.86 & 49.26 \\ 1 & 1.239 & 1.048 & 3.700 \\ 0 & .6798 & .5451 & 3.145 \\ 0 & .5451 & .6720 & 2.714 \end{pmatrix}$$

4Matrix elimination is the necessary solution to the multivariate problem. Appendix 3 contains the computer programs necessary to perform this operation.

52

TABLE 36 Totals for Multivariate Solution

	Soybeans
	Supply Demand

		<u>x_1</u>	<u>x_2</u>	<u>x_1^2</u>	<u>x_2^2</u>	<u>x_1x_2</u>	<u>x_1y</u>	<u>x_2y</u>
	y	(Billions)		(Billions)			(Millions)	
1964	1	2.67	.700	.677	.490	.458	.474	1.869
1965	2	2.88	.845	.738	.714	.545	.624	2.434
1966	3	2.98	.928	.839	.861	.704	.779	2.765
1967	4	2.93	.976	.874	.953	.764	.853	2.860
1968	5	2.69	1.107	.900	1.225	.810	.996	2.978
1969	6	2.63	1.133	.946	1.284	.895	1.072	2.980
1970	7	2.63	1.127	1.230	1.270	1.513	1.386	2.964
1971	8	3.08	1.176	1.258	1.383	1.583	1.479	3.622
1972	9	3.24	1.271	1.202	1.615	1.445	1.528	4.118
1973	10	6.22	1.547	1.283	2.393	1.646	1.985	9.622
1974	11	6.12	1.215	1.435	1.476	2.059	1.744	7.436
1975	12	6.33	1.521	1.200	2.313	1.440	1.825	9.628
Σ		44.40	13.546	12.582	15.977	13.862	14.745	53.276
								49.265

$$\begin{pmatrix} 1 & 0 & .1429 & -1.5240 \\ 0 & 1 & .8018 & 4.6264 \\ 0 & 0 & .2349 & .1922 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & -1.641 \\ 0 & 1 & 0 & 3.9703 \\ 0 & 0 & 1 & .8183 \end{pmatrix}$$

The results show $a = -1.641$, $b = 3.9703$, and $c = .8183$, so that the approximation of the price is

$$\hat{y} = -1.641 + 3.9703x_1 + .8183x_2$$

where x_1 is the production in billions of bushels and x_2 is the price. If the coefficient of supply is the principal factor in the determination of price, and if either x_1 or x_2 been small, it would have indicated that the data to try when determining price components would be a useless answer. In the example, supply-and-demand data would give a price, but perhaps supply and inflation or demand and inflation. To find out which sets of data are best, each combination of data sets and results compared.

Computer programs are available that perform many different number of items to be input as potentially relevant factors. In financial mathematical applications packages. You should not expect increases quickly as more data relationships are included. It is used to find solutions to complex problems of supply and demand to the advantage of not being bound to a solution that fixes the price to a constant value. That is, the multivariate answer to the effect of supply is always four times greater than the univariate solution may say that, in most years when supply is not the primary factor. Neural nets are discussed in Chapter 5.

Selecting Data for an S&P Model

The stock market is driven by a healthy economy and by lower interest rates. A good economy means that there is high employment and consumers are actively buying homes, durable goods, and frivolous items. Low interest rates result in more corporate profitability, which is also influenced by controlled growth and low inflation, both delicate issues orchestrated by the Central Bank.

To create a robust S&P model, whether using multiple regression or, as we will discuss later, neural networks, it is necessary to select the

most meaningful data. The following was suggested by Lincoln to be used for a 6month S&P forecast:⁵

1. S&P price the closing prices of the Standard & Poor's 500 cash index
2. Corporate bonds/Treasury bonds the BAA corporate bond yield divided by the 30year Treasury bond yield, normalized by subtracting the historical mean
3. Annual change in the dollar the 12month change in the dollar, minus 1, which might be based on the Dollar Index traded on the New York Futures Exchange, or a comparable weighting of major currencies
4. Annual change in Federal funds rate the 12month change in the Fed funds rate, minus 1
5. Federal funds rate/discount rate the Fed funds rate divided by the discount rate, normalized by subtracting the historical mean
6. Money supply M1 money supply, not seasonally adjusted
7. Annual Consumer Price Index the 12month change in the CPI, minus 1
8. Inflation/disinflation index the annualized 1month change in the CPI divided by the 12month change in the CPI, minus 1
9. Leading economic indicator the 12month change in the leading economic indicators, minus 1
10. Onemonth versus 10month oscillator for the S&P cash index the difference between the monthly average and the past 10 months, approximately 200 days
11. Inflationadjusted commercial loan the inflationadjusted 12month growth in commercial loans

Lincoln used 20 years of monthly values to forecast the S&P price 6 months ahead. Some of this data is available on a weekly basis and might be adapted to a shorter time frame; however, it is unreasonable to think that an accurate daily forecast is possible using weekly or monthly data. It is also likely that the use of more frequent S&P price data is inconsistent with the monthly statistics and will introduce more noise and make the results less reliable.

One element that is missing from the 11 items listed is a volatility adjustment. There has been a 600% increase in the S&P price over 20 years, and volatility is clearly much higher. A simple percentage relationship may not describe this change adequately, but may be used until a better one is substituted. Therefore, those items that become more or less volatile as prices move higher and lower must be corrected using a volatility normalizing factor. For example, if we consider the initial volatility when the S&P was 100 at the beginning of the data as normal, then when the price is 800 we will divide the current volatility by 8 to normalize, or even take the square root if the increase in volatility is nonlinear. This type of adjustment would apply to nearly all the items except money supply and the inflation/disinflation index. Don't forget that, where applicable, yields must always be used, not prices.

⁵ Thomas H. Lincoln, "Time Series Forecasting: ARMAX," Technical Analysis of Stocks and Commodities (September 1991).

Generalized Multivariate

in general, the relationship between n independent variables is expressed as

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

The solution to this equation is an extension of the problem. $n + 1$ equations in $n + 1$ variables are created by summing the $n + 1$ equations by multiplying the second equation by x_1 , the third by x_2 , and so on, up to the n th equation by x_n . The general equation by multiplying the second equation by x_1 is:

$$\begin{aligned} a_0n + a_1 \sum x_1 + a_2 \sum x_2 + \cdots + a_n \sum x_n &= \sum y \\ a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_1x_2 + \cdots + a_n \sum x_1x_n &= \sum x_1y \\ a_0 \sum x_2 + a_1 \sum x_1x_2 + a_2 \sum x_2^2 + \cdots + a_n \sum x_2x_n &= \sum x_2y \\ &\vdots \\ a_0 \sum x_n + a_1 \sum x_1x_n + a_2 \sum x_2x_n + \cdots + a_n \sum x_n^2 &= \sum x_ny \end{aligned}$$

The solution to this system of equations can be calculated by a standard program available for this purpose (see Appendix 1). One's experience in regression analysis should remember that the solution is only valid within the range of the data points; when projecting outside this range, the predictive qualities of the regression formula decrease. A good solution will be found for interest rates and grain prices at the same levels as previous years, but may not be reliable for projecting erratic new highs over the years.

Many dependent variables (x_i) may be used to increase the correlation. The predictive quality of this solution will increase with independent variables. It is best to start with the obvious factors, such as inflation, then add standard economic statistics, as the trial product, as well as supply-and-demand information, and so on. For both grain and energy markets, the accumulated influence on price; these factors also have an exponential influence represented in terms of an index of adjustment. Measures help determine whether additional factors are necessary.

Least-Squares Sinusoidal

A special case of the multiple linear predictor occurs when the data in the sample time series data. These peaks and valleys have a cyclic pattern. One of the more well-known uses of this method is by Hurst in *The Profit Magic of Stock Transaction Timing* (Prentice-Hall, 1964), an interesting example of Fourier analysis applied to the Dow Jones Industrial Average.

The equation for the approximation of a periodic motion is:

$$y_t = a_0 + a_1t + a_2 \cos \frac{2\pi t}{P} + a_3 \sin \frac{2\pi t}{P} + a_4t \cos \frac{2\pi t}{P} + a_5t \sin \frac{2\pi t}{P}$$

which is a special case of the generalized multivariate approximation:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5$$

where P is the number of data points in each cycle and

$x_1 = t$, the incremental time element

$x_2 = \cos (2\pi t/P)$, a cyclic element

$x_3 = \sin (2\pi t/P)$, a cyclic element

$x_4 = t \cos (2\pi t/P)$, an amplitude-variation element

$x_5 = t \sin (2\pi t/P)$, an amplitude-variation element

The term $a_1 t$ will allow for the linear tendencies an entire cycle and $2\pi t/P$ designates a section ($1/P$) weight to either the sin or cos functions at different

The solution is calculated in the tabular manner of simultaneous linear equations derived in the same way as substituting from the table of sums and solving the complete discussion of curve fitting using trigonometric seasonality").

ARIMA

An *Autoregressive Integrated Moving Average* (ARIMA) is a repeated regression analysis over a moving time window. The ARIMA process automatically applies the most important a preset order and continues to reanalyze results until the best coefficients is found. In Chapter 21, the selection of the best within a trading strategy are discussed. This involves identifying a testing range. An ARIMA model does not require it. Because it is used to recalculate the best fit each time, it can be thought of as the first *adaptive process*.

G.E.P. Box and G.M. Jenkins refined ARIMA at their own procedures for solution have become the industry standard, referred to as the *Box-Jenkins* forecast. The two important concepts are "moving average" and "autoregression." *Autoregression* refers to predicting, for example, using only gold prices to analyze the *average* refers to the normal concept of smoothing the past n days. The moving average and popular Box-Jenkins are discussed in Chapters 4 and 5. This process uses an exponential smoothing, the most popular methods.

In the ARIMA process, the autocorrelation is used to forecast prices will forecast future prices. In a first-order autoregression, the previous day are used to determine the forecast. The

$$P_t = a \times P_{t-1} + e$$

where P_t is the price being forecast (dependent variable)
 P_{t-1} is the price being used to forecast (independent variable)
 a is the coefficient (constant percentage)
 e is the forecast error

In a second-order autoregression, the previous two

$$P_t = a_1 \times P_{t-1} + a_2 \times P_{t-2} + e$$

where the current forecast P_t is based on the two previous forecasts and two unique coefficients and a forecast error. The moving average and forecast error, e . There is also the choice of a first- or second-order

$$\text{First-order: } E_t = e_t - b \times e_{t-1}$$

$$\text{Second-order: } E_t = e_t - b_1 \times e_{t-1} - b_2 \times e_{t-2}$$

where E_t is the approximated error term
 e_t is today's forecast error

⁶ G.E.P. Box and G.M. Jenkins. *Time Series Analysis: Forecasting and Control*. (San Francisco: W. H. Freeman and Co., 1976).

e_{t-1} and e_{t-2} are the two previous forecast errors
 b_1 and b_2 are the two regression coefficients

Because the two constant coefficients, b_1 and b_2 , causing average process is similar to exponential smoothing.

The success of the ARIMA model is determined by autoregression and low variance in the final forecast. To use a first- or second-order autoregression is based on the *coefficients* of the two regressions. If there is little change in the process, it is not used due to its time-consuming nature. It is constructed by adding the moving average term, which makes it an autoregressive process.

$$P'_t = P_t + E_t + e'$$

where P'_t is the new forecast, and e' is the new forecast error. It is again repeated for the new errors e' , added back to the previous error and another error e'' . When the variance of the ARIMA process is complete.

The contribution of Box and Jenkins was to structure the model so that the autoregression and moving average processes are second-order processes. To do this, it was first necessary to make the process *stationary*. Detrending can be accomplished by creating a new series by subtracting each period's value from the previous one. Of course, the ARIMA program must remember all of the values to restore the final forecast to the proper price notation. If a satisfactory solution is not found in the first attempt, the data are still not stationary and further differencing is required.

Because of the three features just discussed, the model is called *ARIMA*(p, d, q), where p is the number of autoregressive terms, d is the number of differences, and q is the number of moving average terms. It is equivalent to simple exponential smoothing, a common forecasting method. In its normal form, the Box-Jenkins ARIMA model is used to determine the best model for a given set of data.

1. *Specification.* Preliminary steps for determining the best model and moving average to be used.

- a. *The variance must be stabilized.* In many cases, the variance is directly related to increased price. In stock prices, the relationship is *log-normal*. A simple test for stationarity, is checked before more complex tests are used.
- b. *Prices are detrended.* This uses the technique of differencing (or more) will be performed to stabilize the properties in the series (this is determined by the variance test).
- c. *Specify the order of the autoregressive and moving average.* This fixes the number of prior terms to be used in the model (usually the same number). In the Box-Jenkins model, the number is small, often one for both. Large numbers require a complex calculation, even for a computer.

How many terms are necessary? This is determined by the variance test and may require manual intervention in this step is to find the fewest terms necessary. Computer programs will print a *correlogram*, a display of the autocorrelation function. The correlogram is used to find whether a

2. Estimation: determining the coefficients. The previous step involves determining the number of autoregressive and moving average terms in the model. The ARIMA method of solution is one of minimum variance. In minimization, it will perform a linear or nonlinear search for the best fit.

FIGURE 39 ARIMA correlograms. (a) Ideal correlogram results, showing significance in the first and seconds lags only. (b) Significance in the first and third lags, indicating that further trend elimination is necessary. (c) No significant correlations at any lag values.

